Natural Language Project Short Paper

Diogo Paixão - ist1113214
Instituto Superior Técnico - Group 68
Lisboa, Portugal

Guilherme Teixeira - ist1113227
Instituto Superior Técnico - Group 68
Lisboa, Portugal

ACM Reference Format:
Diogo Paixão - ist1113214 and Guilherme Teixeira - ist1113227. 2024. Natural Language Project Short Paper. In . ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Given a dataset about movies composed by the title, language, genre, director and plot of the movie, we were tasked to create a machine learning model that could predict the genre of movies.

This paper will give some insights about the strategies followed, models developed and results obtained by the group number 68 in this project.

2 Models

2.1 Data Pre-Processing

The work began by taking a look at the dataset to understand its structure and quality. After loading the data, the key features were analyzed, checked for any missing values and the overall distribution of each variable was explored.

The group discovered that there were 177 entries where the director was listed as "Unknown". Given the relatively small size of the dataset, these samples were not deleted, as removing them could negatively impact the diversity and size of the training data.

The first major adjustment made was removing the "title" column. Although it might seem like a relevant feature, the group thought that the titles themselves did not contain meaningful information for predicting a movie's genre, since they were so random.

For the plot, some essential preprocessing steps were performed. The punctuation and stopwords were removed and the plot's text was converted to lowercase, ensuring consistency. With the text being cleaned up, it was then converted into embeddings, so that it could be fed to machine learning models.

The group decided to experiment this conversion with two different pre-trained models: BERT (bert-base-uncased) and GloVe (with 100 dimensions). The results of using these two methods will be discussed in a later sections of the report.

2.2 Feature Engineering

The group focused on enhancing the original data by creating new features that could provide valuable insights for the predictive models. They are the following:

The plot metrics, including *plot length* (word count), *average word length*, *unique word count*, and *number of named entities* (e.g., people, places, organizations), were created to provide insights about the complexity, linguistic style, and narrative structure.

Sentiment polarity (positive or negative) and *sentiment subjectivity* (degree of factual information in the plot) can give insights about the tone of style of writing.

Lexical diversity indicates the richness of vocabulary that can be connected with the genre.

Part-Of-Speech (POS) indicates the proportion of nouns, verbs and adjectives which can help the model understand how descriptive or action-driven the movie is.

Flesch-Kincaid Readability Score and Latent Dirichlet Allocation were used to find the targeted audience and the two underlying topics about the movies, respectively.

2.3 Models Implemented

The goal was to identify the best approach for this classification task, so, many different machine learning models strategies were explored.

The first model chosen was Logistic Regression (LR), a straightforward model that often provides a strong baseline for text classification problems. This model was the first due to it's simplicity and easiness to implement.

The next model developed was Random Forest (RF), an ensemble method that constructs multiple decision trees and combines their predictions. RF was chosen due to it's strong capability in classification problems.

The group also implemented Support Vector Machines (SVM), which are well-suited for high-dimensional data like text embeddings. SVM's ability to find the optimal decision boundary between classes made it a strong option for this project.

And lastly, the Multi-layer Perceptron (MLP), a type of neural network, was developed. Since MLP can model complex, non-linear relationships in data, the group thought it would be interesting to see how it performed with the embeddings generated from the movie plots, knowing that the quantity of data wasn't big.

After evaluating all four models across both embeddings, the best-performing model was improved using Bagging. involving the training of multiple instances of the best model on different subsets of the data, and combining their predictions. This technique was used to improve the model accuracy and sensitivity.

The following sections will delve into the results of these models and how they performed with the two different types of embeddings.

3 Experimental Setup and Results

The data was split with a ratio of 80/20 (80% to training and 20% do testing) and 20% of the training data was reserved for validation to evaluate the predictions.

Conference'17, July 2017, Washington, DC, USA
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

The group tried to augment the dataset with synonym substitution. However, the accuracy of the models decreased. This is likely due to the synonyms changing the general context of the plot.

To optimize the models, a parameter tuning process was used, driven by pre-defined dictionaries. This was handled through a custom *ModelOptimization* class, which systematically tested different parameter combinations for each algorithm, to find the best one. For example, experimented with different neuron configurations for the MLP.

To evaluate the models, the accuracy was the main metric, which gave an overall idea of how well the models were predicting genres. However, the group also looked at sensitivity (recall), specificity and F1-score.

The numbers 0, 1, 2, 3, 4, 5, 6, 7, 8 on the Confusion Matrix correspond to the genres Action, Animation, Comedy, Crime, Drama, Horror, Romance, Sci-Fi and Western respectively.

Table 1: Results with GloVe Embeddings

Model	Accuracy	Sensitivity	Specificity	F1-Score
LR	0.607	0.595	0.949	0.601
RF	0.574	0.527	0.944	0.554
SVM	0.636	0.603	0.952	0.630
MLP	0.622	0.608	0.951	0.619

Table 2: Results with BERT Embeddings

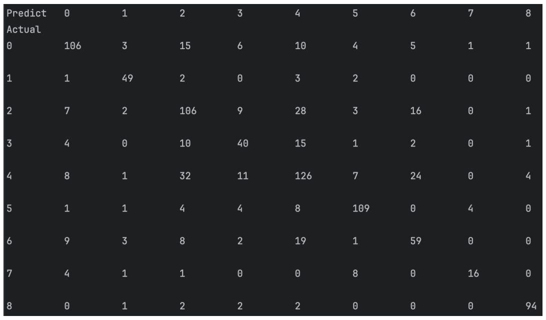
Model	Accuracy	Sensitivity	Specificity	F1-Score
LR	0.674	0.657	0.958	0.666
RF	0.589	0.518	0.946	0.531
SVM	0.676	0.651	0.958	0.683
MLP	0.676	0.664	0.958	0.670
Bagging	0.686	0.669	0.959	0.680

Table 3: Evaluation of the Best Model by Genre

Genre	Accuracy	Sensitivity	Specificity	F1-Score
Action	0.560	0.702	0.961	0.729
Animation	0.766	0.860	0.988	0.831
Comedy	0.434	0.616	0.914	0.602
Crime	0.388	0.548	0.964	0.544
Drama	0.488	0.592	0.896	0.594
Horror	0.772	0.832	0.971	0.820
Romance	0.484	0.584	0.949	0.570
Sci-Fi	0.516	0.533	0.995	0.637
Western	0.895	0.931	0.992	0.931

4 Discussion

When comparing GloVe with BERT embeddings, it's clear that the BERT embeddings are better, as expected. This happens because BERT creates contextual embeddings, considering both right and left words.



```
Predict   0      1      2      3      4      5      6      7      8
Actual
0         106    3      15     6      10     4      5      1      1

1         1      49     2      0      3      2      0      0      0

2         7      2      106    9      28     3      16     0      1

3         4      0      10     40     15     1      2      0      1

4         8      1      32     11     126    7      24     0      4

5         1      1      4      4      8      109    0      4      0

6         9      3      8      2      19     1      59     0      0

7         4      1      1      0      0      8      0      16     0

8         0      1      2      2      2      0      0      0      94
```

Figure 1: Bagging Confusion Matrix

Seeing the results of the automatic evaluation the group reached the conclusion that the MLP model was the best so the bagging was applied to it. MLP was the best, because it's a deep learning model, they can learn complex pattern much better than traditional ML models. The group didn't have many expectation at the beginning, because the small dataset would make a deep learning model prone to overfit, but results showed the opposite. The use of a function that optimizes the parameters of a machine learning model were probably crucial for the success of MLP in such a small dataset.

When looking at the data, the group found that the bigger miss predictions were between drama and comedy. The best model confused both many times. To ensure these suspicions, a confusion matrix was created (Figure 1), which confirmed the speculations.

The movie "Adult world". It's a comedy, but the model predicted drama. Probably because it's about women being kick out of her parent's house and one time she tries to commit suicide.

The movie "The Women". It's a comedy but model thinks it's a drama because it's about a clothing designer who was betrayed by her husband and her friend, while having a lot of familiar issues.

The group thinks that the mix of these heavy themes mixed comic dialogues, makes it hard for even humans to classify this movie.

5 Future Work

While the current feature set has provided valuable insights, a deeper exploration of additional features could bring better results.

Another limitation is the relatively small dataset we used. To address this, we could try to find an oversample method that increases the performance.

Lastly, experiment with more advanced deep learning architectures in the future. Implementing models such as BERT or other deep learning model could improve our results.

6 Conclusion

To conclude, many machine learning models and methods of embedding were tested, and the group concluded that the MLP model using BERT embeddings was the best option tested, and the task to classify movies can be hard. But, we recognize that this project could have been executed better, we could have tried more deep learning models and created better features.

Overall, the experiences gained from this project enriched the group's knowledge about natural language processing.